基于数据立方体挖掘疾病–基因–药物新关联*

魏 星[1,2] 胡德华[1] 易敏寒[1] 朱启贞[1] 朱文婕[2]
[1](中南大学信息安全与大数据研究院 长沙 410083)
[2](蚌埠医学院公共基础学院 蚌埠 233003)

摘要:【目的】在海量文献中，挖掘并预测生物医学实体之间的新关联，构建关联网络。【方法】提出一种基于数据立方体的新方法挖掘疾病–基因–药物间关联，以糖尿病为例，构建关联网络，并使用关联规则量化实体关联程度。【结果】由糖尿病相关疾病(14 种)、基因(23 种)和药物(24 种)构建三个 1-D 方体、三个 2-D 方体及其关联网络和一个 3-D 方体关联网络，共计存在 411 种关联，同时得到 8 个关联子网。【局限】数据预处理存在主观性，可能会对挖掘结果产生影响。【结论】算法性能优于其他同类算法，能够为糖尿病精准医疗提供更好的新研究思路。

关键词: 疾病 基因 药物 数据立方体 关联规则 关联网络
分类号: TP391 G202
DOI: 10.11925/infotech.2096-3467.2017.0641

1 引 言

生物医学文献正在以前所未有的速度增长，其摘要中包含了海量的实验结果、基因表型描述和药效信息，整理挖掘其中有效信息，已成为生物知识发现和生物医学研究中一个重要手段[1]。如何才能有效利用这些文本中所蕴含的生物医学知识，无疑对分析海量生物医学数据是非常重要的，常用方法是通过关键词直接检索，但是这只能从大量文档集合中找到用户需求相关的文件列表，而不能从文本中直接获取用户感兴趣的信息。因此，如何从大规模生物医学文献中自动挖掘相关知识是一项迫在眉睫的任务。常见的生物实体间关联的研究有：蛋白质与基因的关联[2]，药物与药物的关联[3]，药物与疾病的关联[4]等。

数据立方体(Data Cube)[5]能够存放多个数据维(如疾病、基因和药物)上的预计算度量(如关联强度)，用户可以以多维方式，通过如下钻或上卷这样的联机分析处理(OLAP)操作探查数据，进行数据分析和知识发现，探索感兴趣的模式。

本文基于数据立方体探查多维空间中的数据，同时使用关联规则计算实体间的关联度，以糖尿病为例，构建糖尿病相关疾病–基因–药物关联网络，分析并探讨实体间潜在关联，突出并挖掘关联网络中的关键节点，提出实验性研究假设，为研究人员对今后有关糖尿病的诊断与治疗、疾病候选基因筛选、靶向药物和个性化医疗等研究提供数据支持和新的研究思路。

2 相关研究

目前与疾病有关的生物医学文本挖掘研究大多集中在基因的功能信息上，如：对疾病基因和疾病候选基因的分类排序[6]，使用图论构建疾病与疾病基因关联度的网络模型[7]，利用定量性框架模型综合分析疾

通讯作者：胡德华，ORCID: 0000-0001-8027-405X，E-mail: hudehua2000@163.com。

*本文系国家自然科学基金项目“利用黄鳝性逆转模型探索 piRNA 通路在性别决定中的作用机制”(项目编号：31500999)和安徽省高校质量工程“医学院校物联网工程专业建设医工融合的实践教学新模式”(项目编号：2016jyxm0673)的研究成果之一。

病基因与蛋白质之间的作用预测药物新靶点[8]以及计算药物重新定位[9]等，而关于疾病与多个其他实体的关联挖掘属于一个新兴的研究领域。

生物实体关联挖掘方法有多种，如：Lamb 等[10]利用具有生物活性的小分子治疗基因表达谱数据，开发"Connectivity Map"系统，用于挖掘化学与生理过程、疾病与药物之间的小分子共享作用机理，依此挖掘疾病–药物之间的关联。Natarajan[11]在文献中获得疾病、基因的多种特征，从 OMIM 得到已知疾病–基因关联，对比之后，挖掘出 120 对基因–疾病新关联。Odibat 等[12]提出一种基于排序任意重叠定位协同聚类算法，并依此构建判别模型，通过对基因表达数据集的分析运算，可以有效分类疾病基因表达结果。Li 等[13]构建了一个用于判断疾病与候选基因随机集优先级的评分模型，使用基于网络与表型分析的方法在生物医学文献中进行数据挖掘，该模型能够较为精准地将已知致病基因进行排序，同时也能在一定程度上预测新的候选疾病基因。这些研究使用不同方法挖掘生物实体关联，为相关研究提供了多种思路，但使用数据立方体挖掘三个生物实体关联的方法，笔者所知，尚未见报道。CoPub[14]和 PubGene[15]在两者关联挖掘中与本文方法较为类似，但前者挖掘了基因–疾病、药物–疾病的关联，其结果经 ROC 曲线验证后，最高只有 70% (R-scaled Score≥20)，而后者只挖掘了基因–基因间的关联，结果精确度仅有 60%，而且这两项研究并没有将三者关联综合构建网络，分析不够全面。

综上，目前大多数关联挖掘方法都是在两个生物实体之间进行研究的，对三个及三个以上的实体关联挖掘方法研究较少，而且结果精度均不高，这对预测结果的可信度会造成较大影响。因此，本文基于数据立方体将疾病–基因–药物三者结合构建关联网络，挖掘三者之间的新关联，提高算法性能以及挖掘精确度。

3 计算方法及过程

疾病基因药物数据立方体关联网络是由两两关联组合构成的，实现步骤如下：

(1) 对文献进行数据预处理，获得数据立方体的 0-D 顶点方体和三个 1-D 方体；

(2) 设定最小 support 阈值，依据关联规则计算得到三个 2-D 方体内疾病、基因和药物之间的两两关联度；

(3) 使用 BUC 算法构建数据立方体，得到 3-D 基本方体内的实体间关联度；

(4) 利用 R 语言实现多维方体的关联网络的可视化，分析关联网络的分布程度和不同模式的识别程度。

(5) 使用 ROC 曲线验证本文算法的准确性和可靠性。

3.1 数据预处理

由于文献摘要是自然语言书写，属于非结构化数据，所以需先进行数据标准化预处理，不同研究者侧重点不同，本文设定如下步骤进行处理：

(1) 将文献摘要所有字母转为小写；

(2) 把文本转化为单独句子；

(3) 去除标点符号以及与本研究无关的词，如："this"、"an"等；

(4) 替换希腊字母，如："α→Alpha"等；

(5) 基于词典(Gene_Dictionary"和"Drug_dictionary")比对词集中实体名称，若二者与词典中名称(或别名、编号等)相同，即可认定发现了一个实体对象；

(6) 挖掘出所需实体并记录其所在文献的 PMID 号，用于后续关联挖掘。

通过上述算法，将糖尿病相关文献摘要中的基因、药物实体名称进行处理和合并，最终得到规范化的 0-D 方体数据。

3.2 数据立方体

数据立方体(Data Cube)[16]由维和事实定义，维是一个单位(或一次研究)想要记录的透视或实体，常用于商业数据关联挖掘。本文将从 PubMed 中下载的生物医学文献作为数据仓库，创新性地提出将生物实体(如：疾病、基因、药物)作为维，其中每个维都有与之相关联的表，该表称之为维表。同时使用 support、lift 的值作为事实度量标准，这样即将生物实体关联转变为立方体中维与维之间的关联。

在数据立方体中，以 disease、gene、drug 三个属性作为维，以 support 和 lift 的值作为度量，将该立方体计算的方体或分组总数为 8 个，分别为{(disease, gene, drug), (disease, gene), (disease, drug), (gene, drug), (disease), (gene), (drug), ( )}，其中( )意味着分组为空，所以实际上有 7 个分组，这些分组构成了该数据立方体的方格体，如图 1 所示，其中顶点方体(或 0-D 方体)表示分组为空的情况，包含所有可能关联。

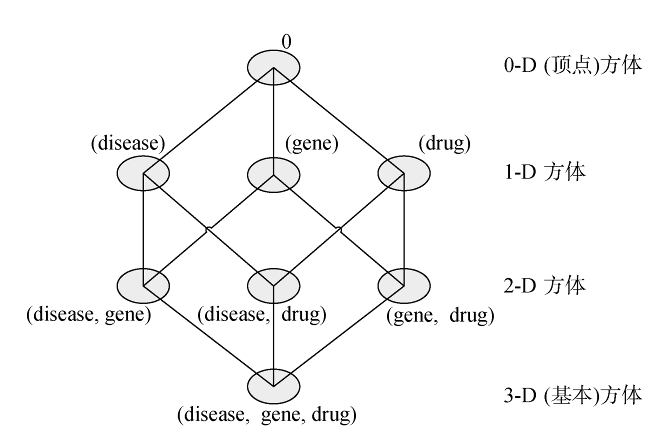


图 1 数据立方

(注: 由方体的格组成三维数据立方体, 每一个立方体代表一个不同的分组; 基本方体包含三个维: disease、gene 和 drug。)

3.3 关联规则

在数据立方体中度量关联时, 需要用到以下术语与度量指标:

(1) 支持度 support 用于衡量集合内各项出现的频次阈值, 如公式(1)所示。

$$\text{support}(A \Rightarrow B) = P(A \cup B) = a / N \qquad (1)$$

(2) 提升指数 lift 能够评估一个预测模型是否有效, 体现集合{A}对{B}的重要性, 如公式(2)所示。

$$\text{lift}(A \Rightarrow B) = \frac{\text{confidence}(A \Rightarrow B)}{P(A)} = \frac{P(A \cup B)}{P(A)P(B)} \qquad (2)$$

若值为 1, 则 A 与 B 无关联; 若值小于 1, 则 A 与 B 相斥; 若值大于 1, 则值越高, A 与 B 之间的关联规则越有价值。本文考虑相关实体可能只是在文献摘要中偶尔或对比提及, 不属于研究内容, 所以设定 life 阈值为 3, 即置信度在 99.8%以上或关键值标准偏差是标准正态分布 3 倍以上, 即认为两者具有强关联性, 如 lift>3 就是具有强关联性。

3.3 BUC 算法

本文的数据立方体实际上是一个稀疏冰山立方体(Iceberg Cube)[17], 因此适合使用自底向上构造(Bottom-Up Construction, BUC)算法[18]构建此数据立方体的关联网络, 该算法自顶向下钻, 即从高聚集单元向较低、更细化的单元移动, 详细算法见文献[17-18]。

3.4 新关联预测

构建实体关联网络后, 可以发现有些关联(即网络中的边)是生物医学资料中从未报道过的新关联, 也就是关联挖掘的假阳性结果, 但这并不意味着这些结果没有用处, 恰恰相反, 这也是生物实体关联挖掘的主要目的之一, 预测新实体关联[19]。这样通过构建三者关联网络再挖掘出的实体新关联(边), 比以往两两实体预测新关联, 具有更高的可信度, 还可能挖掘出更深层次的新关联。最后使用关联规则将所得预测结果计算并排序, 列出可能性最大的实体新关联, 为生物学研究者设计实验方向提供数据支持。

3.5 R 语言实现和 R 曲线验证

R 语言是一种为统计计算和绘图而生的语言和环境, 包含超过 5 000 种开源包(如 igraph 扩展包), 能够较为轻松地构建关联网络[20]。ROC 曲线检测算法的准确性适用于二分类情况, 现已广泛应用于医学诊断实验性能的评价[21]。因此, 本文采用 R 语言实现关联网络, 并用 ROC 曲线判别算法性能。

4 实验过程及结果

4.1 数据来源

从 Entrez GENE[22-23]、Gene Ontology[24]、OMIM[25]、DrugBank[26]等数据库中获取并建立基因和药物标准词典, 命名为"Gene_Dictionary"(共计 40 172 个人类基因词条)和"Drug_Dictionary"(共计 1 763 种药物词条)词典, 词典包括每个基因(药物)的标准名称、别名、同义词、标准编号等属性。以这两个词典为标准进行命名实体识别。

其次, 以糖尿病为例, 在 PubMed 中使用"("diabetes mellitus" [MeSH Terms] OR ("diabetes" [All Fields] AND "mellitus" [All Fields]) OR "diabetes mellitus" [All Fields] OR "diabetes" [All Fields] OR "diabetes insipidus" [MeSH Terms] OR ("diabetes" [All Fields] AND "insipidus" [All Fields]) OR "diabetes insipidus" [All Fields]) AND ("2014/08/20" [PDAT] : "2015/08/20" [PDAT])"为检索策略, 获取一年内与糖尿病相关文献共计 37 373 篇, 并以文本格式保存至本地磁盘。由于本文是对文献的摘要进行实体关联挖掘, 所以剔除其他无用信息(如作者、发表日期等)。

糖尿病分 1 型糖尿病、2 型糖尿病等多种不同病症, 为了深入探讨疾病基因药物之间的关联, 需要对糖尿病进一步分类。在 MeSH 词表中糖尿病属于营养代谢系统疾病和内分泌系统疾病, 分别存在 7 种分类, 糖尿病并发症是相关症状的总称, 如表 1 所示。其中, "Diabetes Mellitus, Type 1"和"Diabetes Mellitus, Type 2"以下简写为"T1DM"和"T2DM"。

表 1 糖尿病在 MeSH 词表中的分类

营养性系统疾病下的分类		内分泌系统疾病下的分类		糖尿病并发症的分类	
英文名称	中文名称	英文名称	中文名称	英文名称	中文名称
Diabetes Mellitus, Experimental	实验性糖尿病	Diabetes Complications	糖尿病并发症	Diabetic Angiopathies	糖尿病性血管病
Diabetes Mellitus, Type 1	1 型糖尿病	Diabetes, Gestational	妊娠糖尿病	Diabetic Cardiomyopathies	糖尿病性心肌病
Diabetes Mellitus, Type 2	2 型糖尿病	Diabetes Mellitus, Experimental	实验性糖尿病	Diabetic Coma	糖尿病性昏迷
Diabetes, Gestational	妊娠糖尿病	Diabetes Mellitus, Type 1	1 型糖尿病	Diabetic Ketoacidosis	糖尿病性酮症酸中毒
Diabetic Ketoacidosis	糖尿病酮症酸中毒	Diabetes Mellitus, Type 2	2 型糖尿病,	Diabetic Nephropathies	糖尿病性肾病
Donohue Syndrome	多诺霍综合症	Donohue Syndrome	多诺霍综合症	Diabetic Neuropathies	糖尿病性神经病
Prediabetic State	糖尿病前期	Prediabetic State	糖尿病前期	Fetal Macrosomia	巨大胎儿(症)

4.2 0-D 顶点方体

本文的 0-D 顶点方体，即预处理后得到的“(all)词项集”，是糖尿病数据立方体的顶点，也是后续研究的数据基础。

4.3 1-D 方体疾病维、基因维和药物维

综合表 1，去重后得到：实验性糖尿病、1 型糖尿病、糖尿病性血管病、糖尿病性昏迷等共计 14 种糖尿病相关病症，由此构建数据立方体中 1-D 方体(disease)维；以“Gene_Dictionary”词典为标准，对糖尿病数据立方体中的 0-D 顶点立方体进行过滤，由于可能部分基因在摘要中只是偶尔提及，为了排除干扰，设定 support 阈值为 0.1%，得到 ABCC8 等 23 种基因的 support 值满足大于最小支持度(≥0.1%)的条件，由此构建 1-D 方体(gene)维；以“Drug_Dictionary”词典为标准，对糖尿病数据立方体中的 0-D 顶点方体进行过滤，设定 support 阈值为 0.1%，得到三磷酸腺苷等 24 种药物的 support 值满足大于最小支持度(≥0.1%)的条件，由此构建 1-D 方体(drug)维。

4.4 2-D 疾病基因方体关联网络

依据前述关联算法，得到 14 种糖尿病相关病症和 23 种基因产生的 194 种关联，其中 2 型糖尿病、糖尿病性神经病、糖尿病性肾病和实验性糖尿病与 23 种基因均具有关联；1 型糖尿病不与 IPF1 和 SUMO4 关联，与其他 21 种基因相关；糖尿病性心肌病不与基因 GAD2、IPF1 和 SUMO4 关联；糖尿病性血管病不与基因 DAD2、IPF1、PTPRN 和 SUMO4 相关；与妊娠糖尿病、糖尿病酮症酸中毒、糖尿病性昏迷相关的基因分别有 11、8、2 种。由此得到 2-D(disease, gene)方体，并生成糖尿病相关疾病基因关联网络，如图 2 所示。

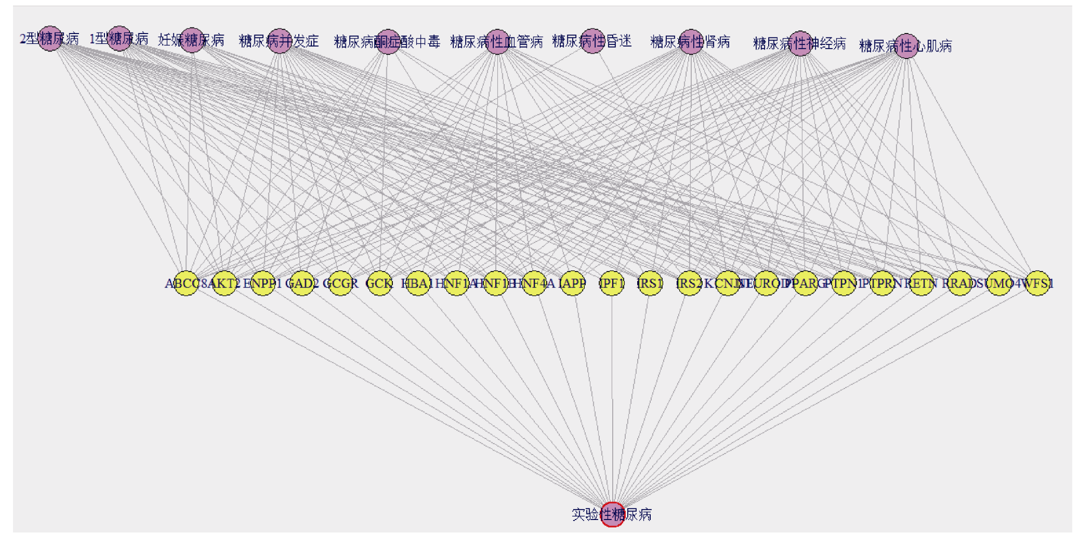


图 2 (disease, gene)2-D 方体的关联网络

图 2 中的实验性糖尿病、2 型糖尿病、糖尿病性肾病和糖尿病性神经病这 4 种糖尿病相关病症与 23 种基因均具有关联。1 型糖尿病不与基因 IPF1、SUMO4 相关，糖尿病性血管病不与基因 GAD2、IPF1、PTPRN、SUMO4 相关，糖尿病并发症不与基因 SUMO4、WFS1 相关，糖尿病性心肌病不与基因 GAD2、IPF1、SUMO4 相关，但这 4 种病症与剩下的其他基因具有关联性。多诺霍综合症、糖尿病前期和巨大胎儿(症)与本文得到的 23 种基因均不具有关联性。

4.5 2-D 疾病药物方体关联网络

通过关联算法计算，有 10 种药物与糖尿病相关病症无关，分别是：β-D-葡萄糖(Beta-D-Glucose)、糖类多酮类复合化合物 19(Compound 19)、布洛芬(Ibuprofen)等；有 4 种病症与药物之间不存在关联，分别是：糖尿病性昏迷、巨大胎儿(症)、糖尿病酮症酸中毒和多诺霍综合症。最终得到 24 种药物和 11 种糖尿病相关病症，以及它们之间的 75 种关联，由此生成 2-D(disease, drug)方体，使用 R 语言构建该关联网络，如图 3 所示。

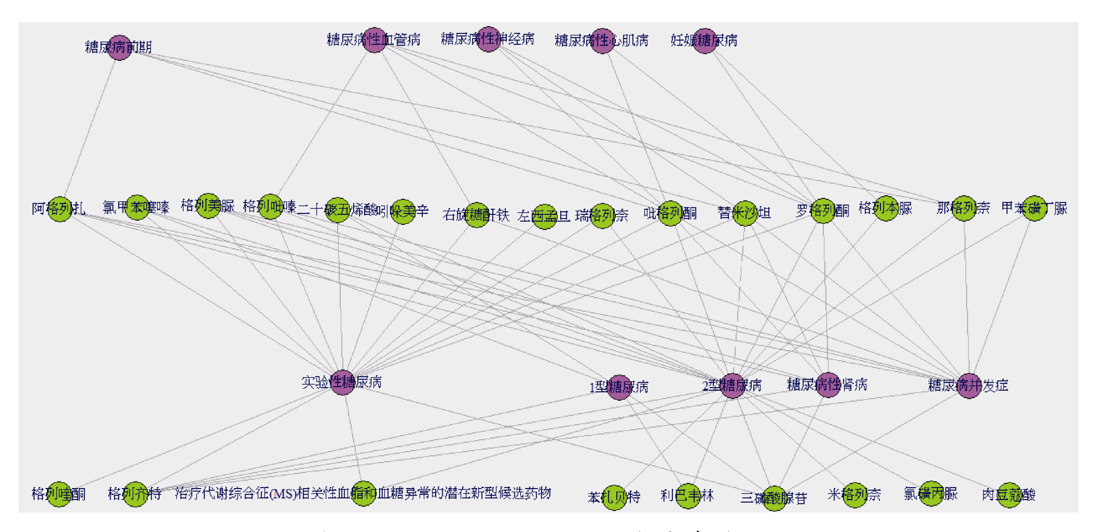


图 3 (disease, drug)2-D 方体的关联网络

4.6 2-D 基因药物方体关联网络

通过关联算法，得到 14 种基因与 15 种药物组成的 142 种关联，构成了 2-D(gene, drug)方体，关联网络如图 4 所示，得到基因 WFS1 与药物三磷酸腺苷具有单相关性，而其他的关联均具有多重性，即一种基因关联多种药物或一种药物关联多种基因。如：基因 ABCC8 对应 10 种药物，药物二十碳五烯酸对应 8 种基因。ATP 敏感性钾通道中产生的变体 E23K 和 S1369A 可以在基因 ABCC8、KCNJ11 中找到，这 2 个变体可能会对一些药物，如：那格列奈等，在治疗 2 型糖尿病的过程中产生抑制力[27]。

4.7 3-D 疾病基因药物方体关联网络

使用 BUC 算法构建糖尿病基因药物数据立方体，设定最小 lift 阈值为 3，去重后，14 种糖尿病病症、23 种基因和 24 种药物之间得到 411 种关联，使用 R 语言构建出糖尿病数据立方体的(disease, gene, drug)3-D 基本方体的关联网络，如图 5 所示。

同时，为了深入探讨每种糖尿病病症的疾病基因药物之间的关联，分别对 8 种糖尿病症状构建子网模型，如图 6 所示。

相关研究发现，苯扎贝特与 2 型糖尿病有较大关联，对于治疗 2 型糖尿病具有较好的疗效，有助于血糖调节[28]；同样，格列喹酮在实验性糖尿病中的治疗效果，也有相关文献[29]进行过报道；2 型糖尿病的易感基因 KCNJ11 部分发病机制也得到验证[30]。

由图 5 和图 6 可得，糖尿病疾病、基因和药物这三者都均有关联性的有 318 组，如：(1 型糖尿病，ABCC8，三磷酸腺苷)，(2 型糖尿病，ABCC8，苯扎贝特)，(实验性糖尿病，ENPP1，瑞格列奈)等。从疾病角度分析，有 9 种糖尿病病症存在三者关联，如：1 型糖尿病有 19 组，32 种两两关联；2 型糖尿病有 126 组，153 种关联等，其中妊娠糖尿病组数最少，只有 8 组 15 种关联，与 5 种基因和 2 种药物之间存在关联。在糖尿病并发症中，糖尿病性肾病组数最多，有 60 组及 80 种两两关联，属于关联网络中的关键节点。

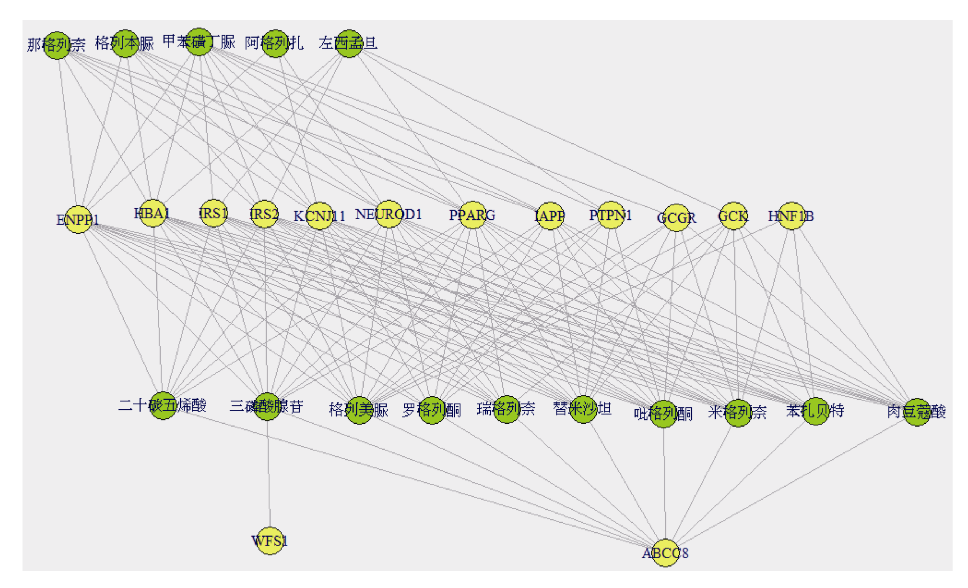


图 4　(gene, drug)2-D 方体关联网络

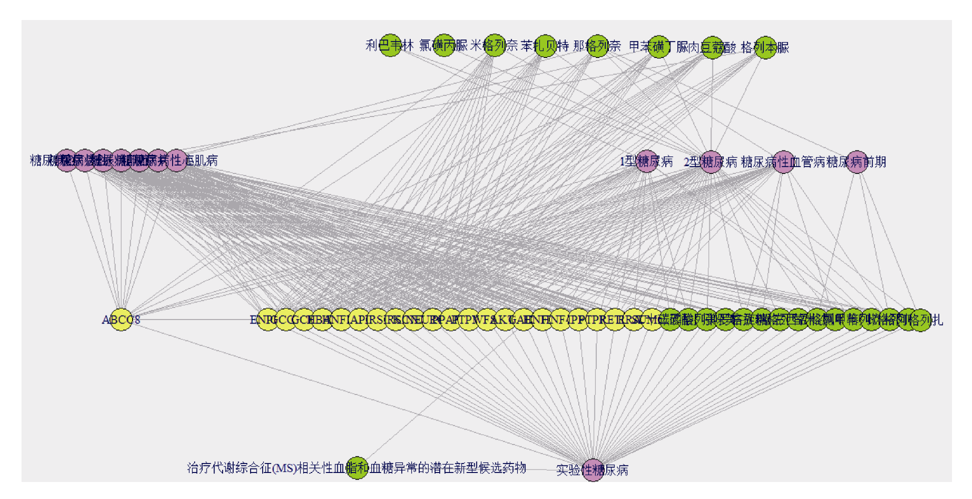


图 5　(disease, gene, drug)3-D 基本方体关联网络

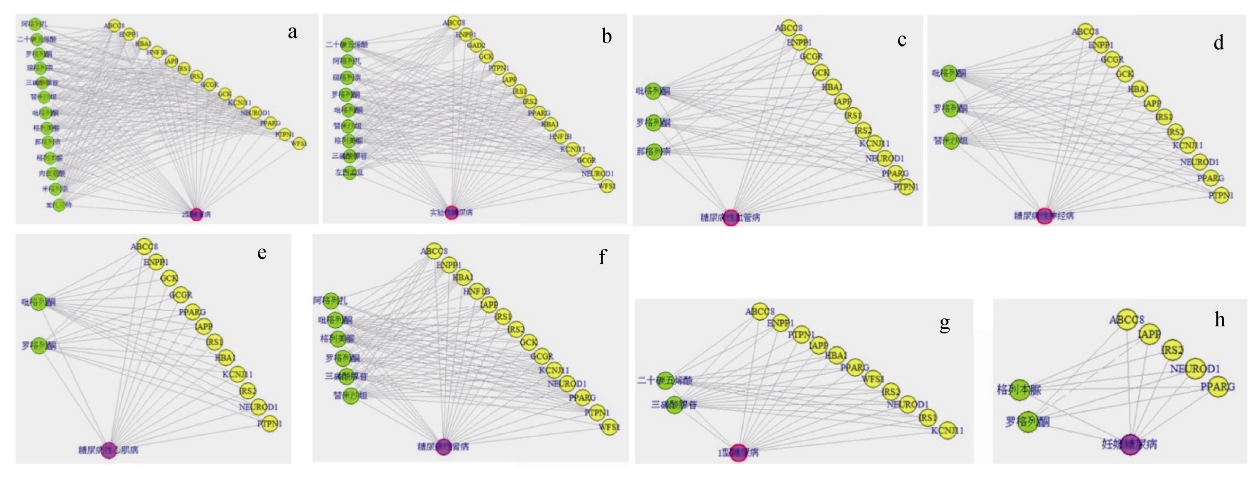


图 6　8 种疾病关联子网

(注: a: 2 型糖尿病; b: 实验性糖尿病; c: 糖尿病血管病; d: 糖尿病性神经病; e: 糖尿病心肌病; f: 糖尿病肾病; g: 1 型糖尿病; h: 妊娠糖尿病)

本文将整个关联网络以糖尿病相关病症为标准，分解出 8 个子关联网络，更有助于发现糖尿病相关病症中的候选基因和候选药物，以及推断疾病、基因药物间的新关联。例如：对疾病基因 2-D 方体的关联网络研究发现，基因 ABCC8 和 KCNJ11 与 2 型糖尿病具有相关性，这两种基因的变异可以引起新生儿童糖尿病以及家族性持续性高胰岛素低血糖症[31]，但在不同人种中的实验结果存在差异。

4.8 新关联预测结果

部分糖尿病的疾病–基因、疾病–药物和基因–药物之间关联程度排名靠前但尚未报道的实体对如表 2 所示，其中基因名参照《英汉人类基因词典》[32]。

表 2 预测部分关联程度较高但尚未证实的生物实体间新关联

Rel	EN 1	Description 1	EN 2	Description 2
Disease-Gene	Diabetic Neuropathies	糖尿病性神经病	IPF1	transcription factor 1
	Diabetic Neuropathies	糖尿病性神经病	SUMO4	small ubiquitin-like modifier 4
	Diabetic Nephropathies	糖尿病性肾病	IPF1	transcription factor 1
	Diabetic Nephropathies	糖尿病性肾病	SUMO4	small ubiquitin-like modifier 4
Disease-Drug	Iron Dextran	右旋糖酐铁	Diabetic Angiopathies	糖尿病性血管病
	GFT505	治疗代谢综合征(MS)相关性血脂和血糖异常的潜在新型候选药物	T2DM	2 型糖尿病
	Telmisartan	替米沙坦	Diabetic Neuropathies	糖尿病性神经病
	Aleglitazar	阿格列扎	Diabetic Nephropathies	糖尿病性肾病
Gene-Drug	IRS2	insulin receptor substrate 2	Icosapent	二十碳五烯酸
	PPARG	peroxisome proliferator-activated receptor gamma	Icosapent	二十碳五烯酸
	IRS2	insulin receptor substrate 2	Levosimendan	左西孟旦
	GCK	glucokinase (hexokinase 4)	Levosimendan	左西孟旦
	ENPP1	ectonucleotide pyrophosphatase/ phosphodiesterase 1	Myristic Acid	肉豆蔻酸

(注: EN as Entity_Name. Rel as Relation)

表 2 中尚未证实的成对关联，可为研究人员提供新的研究思路，例如：目前尚无文献报道基因 SUMO4 与糖尿病性神经病、糖尿病性肾病之间是否存在关联，不过，文献[33]指出，1 型糖尿病患者中的 SUMO4 基因多态性 M55V 与糖尿病性视网膜病变的患病率降低有关，认为通过 SUMO4 蛋白质转译后的修改可能导致某些糖尿病并发症的发展，它们之间存在关联的可能性较大。有报道[34]称某位患者体内的基因 ABCC8 的 34 号外显子突变，导致新生儿肾病，但是由于该患者开始时被误诊为 1 型糖尿病，从而错过了最佳治疗时间，最终发展为肾病晚期，这也间接证明了基因 ABCC8 与糖尿病性肾病可能具有相关性。

4.9 ROC 曲线评价

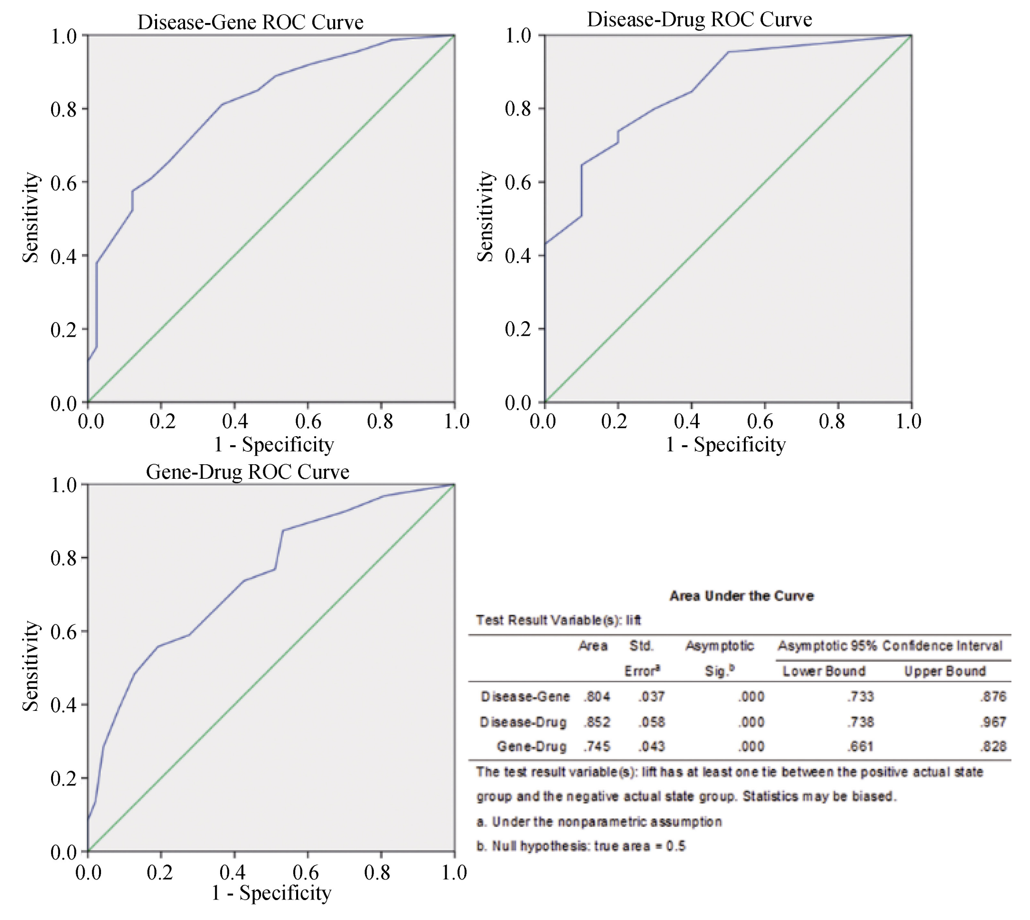
对本文得到糖尿病的疾病–基因、疾病–药物和基因–药物之间的所有关联结果进行准确性验证，关联验证标准[14]如下：

(1) 真阳性(TP)：有已知且确定的直接关联或共现次数大于等于 3，例如：2 型糖尿病与基因 ABCC8[35]；

(2) 假阳性(FP)：无直接关联且共现次数小于 3。

在 SPSS20 环境下使用 ROC 曲线判断算法性能，如图 7 所示。ROC 曲线下的面积分别为 0.804、0.815 和 0.745，关联准确度中等偏上，相应的标准误分别为 0.037、0.076 和 0.043，P 值均为 0.000，95%置信区间分别为(0.733, 0.876)、(0.666, 0.964)和(0.661, 0.828)。

与其他关联挖掘算法[36-37]类似，本文也得到一些假阳性(预测性)结果，这也是生物医学实体关联挖掘的目标之一：提出预测性的研究假设，帮助科研人员设计相关实验方向[38]。



Area Under the Curve

Test Result Variable(s): lift

	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
Disease-Gene	.804	.037	.000	.733	.876
Disease-Drug	.852	.058	.000	.738	.967
Gene-Drug	.745	.043	.000	.661	.828

The test result variable(s): lift has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

图 7 ROC 曲线性能评价

5 结 语

空间数据立方体概念建模需要定义两种元数据：一是来自多种数据源、可维护的、集成的、具有模型数据结构的仓库元数据；二是可以满足决策者分析需求的、集成的仓库元数据[39]。本文定义识别得到的实体(如疾病、基因、药物)为第一种元数据，而文献资料和词典则定义为第二种元数据，因此，得到一种基于网络的多维数据集模型。

此外，本文并没有检索到一些糖尿病常用药物：如胰岛素(Insulin)和二甲双胍(Metformin)等，原因如下：本算法检索时使用的药物词典是 FDA 公布的 Drug_Bank 数据库，其中胰岛素有 9 种名称：(Insulin Regular)、(Insulin Glargine)、(Insulin Lispro)、(Insulin, Porcine)、(Inhaled insulin)、(Insulin Aspart)、(Insulin Detemir)、(Insulin Glulisine)和(Insulin, Isophane)，在文献摘要中完全匹配的检索结果均为零；而二甲双胍在这一年内的糖尿病相关文献摘要中，只检索到 10 篇(support 值约为 0.026%)，小于设定的 support 阈值(=0.1%)。

本研究扩展了网络模式分析疾病–药物–基因关联，网络中的节点代表生物医学实体存储在 RDF 三元组(即疾病、药物、基因)，边表示生物医学实体间的关联(如“谓词”关系)。为简单起见，关联均设为单向关联，丢弃了边的方向和类型，即只要两节点间有关联，便认为这两个节点间有边。这样简化疾病药物基因的关联网络中的网络模式有两点作用：基本可以代表疾病基因药物之间的相互关系；反映了一个可以有效实现特定功能的框架。

本文创新在于：在生物实体关联挖掘研究领域，提出一种基于数据立方体的新方法，挖掘实体关联，并结合关联规则对实体关联程度进行分析排序；以疾病–基因–药物这三种不同生物实体为研究对象，挖掘新关联，而 CoPub[14]挖掘的是基因–疾病、药物–疾病的

关联, PubGene[15]仅挖掘基因–基因间的关联, Sun 等[36]挖掘药物–药物间的关联; 使用 ROC 曲线验证本文算法得到曲线下面积分别为 0.804、0.815 和 0.745, 优于同类算法(如: CoPub 和 PubGene), 因此本文算法性能更高。下一步工作是在更大规模数据中评估本算法的性能, 确保推广效果。

参考文献:

[1] Moreau Y, Tranchevent L C. Computational Tools for Prioritizing Candidate Genes: Boosting Disease Gene Discovery[J]. Nature Reviews Genetics, 2012, 13(8): 523-536.

[2] Fundel K, Kuffner R R. RelEx——Relation Extraction Using Dependency Parse Trees[J]. Bioinformatics, 2007, 23(3): 365-371.

[3] Bui Q C, Sloot P M, van Mulligen E M, et al. A Novel Feature-Based Approach to Extract Drug-Drug Interactions from Biomedical Text[J]. Bioinformatics, 2014, 30(23): 3365-3371.

[4] Xu R, Wang Q Q. Large-scale Extraction of Accurate Drug-Disease Treatment Pairs from Biomedical Literature for Drug Repurposing[J]. BMC Bioinformatics, 2013, 14(13): 1-11.

[5] Gray J, Bosworth A, Layman A, et al. Data Cube. A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Total[J]. Data Mining & Knowledge Discovery, 1997, 1(1): 29-53.

[6] Piro R M. Computational Approaches to Disease-Gene Prediction: Rationale, Classification and Successes[J]. Febs Journal, 2012, 279(5): 678-696.

[7] Goh K I, Cusick M E, Valle D, et al. The Human Disease Network[J]. Proceedings of the National Academy of Sciences of the United States of America, 2007, 104(21): 8685-8690.

[8] Suthram S. Network-Based Elucidation of Human Disease Similarities Reveals Common Functional Modules Enriched for Pluripotent Drug Targets[J]. PLoS Computational Biology, 2010, 6(2): e1000662.

[9] Arrell D K, Terzic A. Network Systems Biology for Drug Discovery[J]. Clinical Pharmacology & Therapeutics, 2010, 88(1): 120-125.

[10] Lamb J, Craeford E D, Peck D, et al. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease[J]. Science, 2006, 313(5795): 1929-1935.

[11] Natarajan N. Inductive Matrix Completion for Predicting Gene-Disease Associations[J]. Bioinformatics, 2014, 30(12): 60-68.

[12] Odibat O, Reddy C K. Efficient Mining of Discriminative Co-clusters from Gene Expression Data[J]. Knowledge & Information Systems, 2014, 41(3): 667-696.

[13] Li J, Edwards S M, Bo T, et al. A Random Set Scoring Model for Prioritization of Disease Candidate Genes Using Protein Complexes and Data-Mining of GeneRIF, OMIM and PubMed Records[J]. BMC Bioinformatics, 2014, 15(22): 3946-3959.

[14] Frijters R, Vugt M V, Smeets R, et al. Literature Mining for the Discovery of Hidden Connections Between Drugs, Genes and Diseases[J]. PLoS Computational Biology, 2010, 6(9): e10000943.

[15] Jenssen T K, Laegreid A, Komorowski J, et al. A Literature Network of Human Genes for High-Throughput Analysis of Gene Expression[J]. Nature Genetics, 2001, 28(1): 21-28.

[16] Li C, Ooi B C, Tung A K H, et al. DADA: A Data Cube for Dominant Relationship Analysis[C]// Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data. 2006: 659-670.

[17] Fang M, Shivakumar N, Garcia-Molina H, et al. Computing Iceberg Queries Efficiently[C]// Proceedings of the 24th International Conference on Very Large Data Bases. 1998: 299-310.

[18] Beyer K S, Ramakrishnan R. Bottom-Up Computation of Sparse and Iceberg CUBEs[C]// Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data. 1999.

[19] Gonzalez G H, Tahsin T, Goodale B C, et al. Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery[J]. Briefings in Bioinformatics, 2016, 17(1): 33-42.

[20] Development Core R Team. R: A Language and Environment for Statistical Computing[J]. Computing, 2013, 14: 12-21.

[21] Hanley J A, Mcneil B J. The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve[J]. Radiology, 1982, 143(1): 29-36.

[22] Donna M, Jim O, Pruitt K D, et al. Entrez Gene: Gene-Centered Information at NCBI [J]. Nucleic Acids Research, 2007, 39(2): 54-58.

[23] Pruitt K D, Tatiana T, Maglott D R. NCBI Reference Sequences (RefSeq): A Curated Non-Redundant Sequence Database of Genomes Transcripts and Proteins[J]. Nucleic

Acids Research, 2008, 33: 501-504.
[24] Ashburner M, Ball C A, Blake J A, et al. Gene Ontology: Tool for the Unification of Biology[J]. Nature Genetics, 2000, 25(1): 25-29.
[25] Hamosh A, Scott A F, Amberger J S, et al. Online Mendelian Inheritance in Man (OMIM), A Knowledgebase of Human Genes and Genetic Disorders[J]. Nucleic Acids Research, 2005, 33(1): 514-517.
[26] Knox C, Law V, Jewison T, et al. DrugBank 3.0: A Comprehensive Resource for 'Omics' Research on Drugs[J]. Nucleic Acids Research, 2011, 39(S1): 1035-1041.
[27] Lang V Y, Fatehi M, Light P E. Pharmacogenomic Analysis of ATP-Sensitive Potassium Channels Coexpressing the Common Type 2 Diabetes Risk Variants E23K and S1369A [J]. Pharmacogenetics & Genomics, 2012, 22(3): 206-214.
[28] Tenenbaum A, Fisman E Z. Balanced Pan-PPAR Activator Bezafibrate in Combination with Statin: Comprehensive Lipids Control and Diabetes Prevention?[J]. Cardiovascular Diabetology, 2012, 11(2): 140.
[29] Ke J T, Li M, Xu S Q, et al. Gliquidone Decreases Urinary Protein by Promoting Tubular Reabsorption in Diabetic Goto-Kakizaki Rats[J]. Journal of Endocrinology, 2014, 220(2): 129-141.
[30] Hui Z, Min G, Zhou T, et al. An Isogenic Human ESC Platform for Functional Evaluation of Genome-wide-Association-Study-Identified Diabetes Genes and Drug Discovery[J]. Cell Stem Cell, 2016, 9: 326-340.
[31] Nichols C G, Koster J C, Remedi M S. Beta-cell Hyperexcitability: From Hyperinsulinism to Diabetes[J]. Diabetes Obesity & Metabolism, 2007, 9 (S2): 81-88.
[32] 张闻. 英汉人类基因词典[M]. 北京: 人民卫生出版社, 2011. (Zhang Wen. English Chinese Dictionary of Human Genes [M]. Beijing: People's Medical Publishing House, 2011.)
[33] Rudofsky G, Schlotterer A, Humpert P M, et al. A M55V Polymorphism in the SUMO4 Gene is Associated with a Reduced Prevalence of Diabetic Retinopathy in Patients with Type 1 Diabetes[J]. Experimental & Clinical Endocrinology & Diabetes, 2007, 116(1): 14-17.
[34] Esmatjes E, Jimenez A, Diaz G, et al. Neonatal Diabetes with End-Stage Nephropathy Pancreas Transplantation Decision[J]. Diabetes Care, 2008, 31(11): 2116-2117.
[35] Stefanski A, Majkowska L, Ciechanowicz A, et al. The Common C49620T Polymorphism in the Sulfonylurea Receptor Gene (ABCC8), Pancreatic Beta Cell Function and Long-Term Diabetic Complications in Obese Patients with Long-Lasting Type 2 Diabetes Mellitus[J]. Experimental & Clinical Endocrinology & Diabetes, 2007, 115(5): 317-321.
[36] Sun K, Liu H, Yeganova L, et al. Extracting Drug-Drug Interactions from Literature Using a Rich Feature-Based Linear Kernel Approach [J]. Journal of Biomedical Informatics, 2015, 55: 23-30.
[37] Rong X, Wang Q Q. Large-scale Automatic Extraction of Side Effects Associated with Targeted Anticancer Drugs from Full-Text Oncological Articles[J]. Journal of Biomedical Informatics, 2015, 55: 64-72.
[38] Gonzalez G H, Tahsin T, Goodale B C, et al. Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery[J]. Briefings in Bioinformatics, 2015, 29: 1-10.
[39] Boulil K, Bimonte S, Pinet F. Conceptual Model for Spatial Data Cubes: A UML Profile and Its Automatic Implementation[J]. Computer Standards & Interfaces, 2014, 38: 113-132.

作者贡献声明:

魏星: 研究方法设计与实现, 论文撰写、修改以及最终版本修订;
胡德华: 提出总体研究思路, 论文修改;
易敏寒: 医学用语修订;
朱启贞, 朱文婕: 算法实现与验证。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: weixing911119@163.com。
[1] 魏星. DM_pubmed_result.txt. 数据立方体 0-D 维.
[2] 魏星. 1-D 维.xlsx. 数据立方体 1-D 维.
[3] 魏星. DM_DiseaseGeneDrug.xlsx. 生物实体关联数据.
[4] 魏星. DISEASE-GENE.sav. 疾病–基因 ROC 曲线验证.
[5] 魏星. DISEASE-Drug.sav. 疾病–药物 ROC 曲线验证.
[6] 魏星. GENE-Drug.sav. 基因–药物 ROC 曲线验证.

收稿日期: 2017-07-03
收修改稿日期: 2017-07-28

Extracting Disease-Gene-Drug Correlations Based on Data Cube

Wei Xing[1, 2] Hu Dehua[1] Yi Minhan[1] Zhu Qizhen[1] Zhu Wenjie[2]
[1](Institute of Information Security and Big Data, Central South University, Changsha 410083, China)
[2](School of Basic Courses, Bengbu Medical College, Bengbu 233003, China)

Abstract: [**Objective**] This study aims to construct a disease-gene-drug correlation network for diabetes mellitus (DM). [**Methods**] First, we proposed a new data cube-based approach to construct a disease-gene-drug correlations network for the DM. Then, we measured the associations among the biological entities. [**Results**] We retrieved the needed data from the PubMed database and constructed three 1-D vertex cubes, three 2-D square cubes and one 3-D disease-gene-drug network, which revealed 411 associations among the 14 subclasses of DM, 23 genes, and 24 drugs. We also constructed 8 optimal disease-gene-drug subnetworks of DM. [**Limitations**] There were some subjective issues with the data analysis. The changing of user behaviors may also influence the results. [**Conclusions**] The proposed algorithm is better than the existing ones, which provides new directions for research on customized medical treatments.

Keywords: Disease Gene Drug Data Cube Association Rules Correlations Network

微软携手亚马逊推出全新 Gluon 深度学习库

据外媒报道，近日微软与亚马逊宣布正式达成战略合作，并联手推出全新深度学习库“Gluon”。届时，Gluon 接口将为开发者们提供一个 Python API 和预先构建的神经网络组件，让他们可以更加流畅地调试和更新。当前，该深度学习库仅支持 Apache MXNet。不过微软表示，将很快支持该公司的认知工具包(CNTK)。构造一个神经网络的难题，在于保持模型构建和训练性能之间的平衡。以 Apache MXNet 深度学习引擎为例，从开发者的角度来看，微软认知工具包(Microsoft Cognitive Toolkit)和 Google TensorFlow 确实可以在一定程度上优化训练的过程，但通常需要大量的时间和复杂的编码。而 Gluon，则为开发者们提供了针对各种神经网络模型的试验接口，以及对底层性能几乎没有任何影响的训练方法。

微软人工智能研究执行副总裁 Eric Boyd 表示，Gluon 接口可以给开发者们“相当自由的选择”。至于它能对整个机器学习社区发挥多大的影响力，仍有待时间去检验。

(来自: http://www.afenxi.com/post/48391)

(本刊讯)